“HOT” ChatGPT: The Promise of ChatGPT in Detecting and Discriminating Hateful, Offensive, and Toxic Comments on Social Media

Year: 2023 | *Citations: 122* | *Authors: Lingyao Li, Lizhou Fan, Shubham Atreja*

Abstract

Harmful textual content is pervasive on social media, poisoning online communities and negatively impacting participation. A common approach to this issue is developing detection models that rely on human annotations. However, the tasks required to build such models expose annotators to harmful and offensive content and may require significant time and cost to complete. Generative AI models have the potential to understand and detect harmful textual content. We used ChatGPT to investigate this potential and compared its performance with MTurker annotations for three frequently discussed concepts related to harmful textual content on social media: Hateful, Offensive, and Toxic (HOT). We designed five prompts to interact with ChatGPT and conducted four experiments eliciting HOT classifications. Our results show that ChatGPT can achieve an accuracy of approximately 80% when compared to MTurker annotations. Specifically, the model displays a more consistent classification for non-HOT comments than HOT comments compared to human annotations. Our findings also suggest that ChatGPT classifications align with the provided HOT definitions. However, ChatGPT classifies “hateful” and “offensive” as subsets of “toxic.” Moreover, the choice of prompts used to interact with ChatGPT impacts its performance. Based on these insights, our study provides several meaningful implications for employing ChatGPT to detect HOT content, particularly regarding the reliability and consistency of its performance, its understanding and reasoning of the HOT concept, and the impact of prompts on its performance. Overall, our study provides guidance on the potential of using generative AI models for moderating large volumes of user-generated textual content on social media.